LUTTE CONTRE LE PHISHING

Revue de Littérature sur la Détection de Phishing

Armand Florent TSAFACK PIUGIE

TEICEE ET UNICAEN

Laboratoire GREYC

ftsafack@teicee.com

armand-florent.tsafack-piugie@unicaen.fr

GREYC
Electronics and Computer Science Laboratory

Normandie Université

UNICAEN

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

cnrs

téicée®
la raison de votre confiance

Qui suis-je ?

Nom : Tsafack Piugie

Prénom : Armand Florent

Adresses Mails: armand-florent.tsafack-piugie@unicaen.fr | ftsafack@teicee.com

Tel : 06 60 44 14 65

Parcours Académique

- **2015-2016 :** Baccalauréat Scientifique
- **2016-2019 :** Licence en Physique option EEA
- **2019-2021 :** Master of Sciences

Parcours Professionnel

- **2021-2022 :** Compétences en IA, Enseignement et chargé de TD
- **2022-2023 :** Cours Réseau Neurones M2 Pro, Direction R&D et ML engineer à Agrix-Tech
- **Mars 2024 :** Doctorant CIFRE 1ère année (**ANRT - téïcée - GREYC** de l'**UCN**) en cybersécurité travaillant sur ***la Lutte contre le Phishing en utilisant l'IA*.**
- Au **GREYC**, je suis à l'équipe ***SAFE*** *(**S**écurité **A**rchitecture **F**orensique biom**E**trie)*

CIFRE : ANRT- téïcée - GREYC

téïcée
la raison de votre confiance

GREYC
Electronics and Computer Science Laboratory

Encadrement de la thèse

GREYC – Equipe SAFE

- Directeur de thèse : **Emmanuel Giguet**
- Co-directeur : **Christophe Rosenberger**

téïcée

- Responsable pôle R&D : **Mathieu Valois**
- Gérant : **Philippe Chauvat**

Plan

1. Contexte
2. Problème
3. Objectifs : Général et spécifiques
4. Enjeux Scientifiques de la recherche
5. Plan de travail
6. État de l’art
7. Mes travaux durant ce premier semestre
8. Formations SYGAL suivies

1. Contexte

➢ La messagerie électronique

- En 2023, **121 Billions de mails échangés** entre **4.3 Billions de personnes** *[XoMEDIA,2024]*
- **~1.4 Milliards** de mails par jour sont envoyés en France *[Corthesy,2022]*
- ***53% du trafic mondial des emails*** sont des Spams et causent des pertes d'environ **20M USD**.
- Synoptique d'envoi d'un e-mail

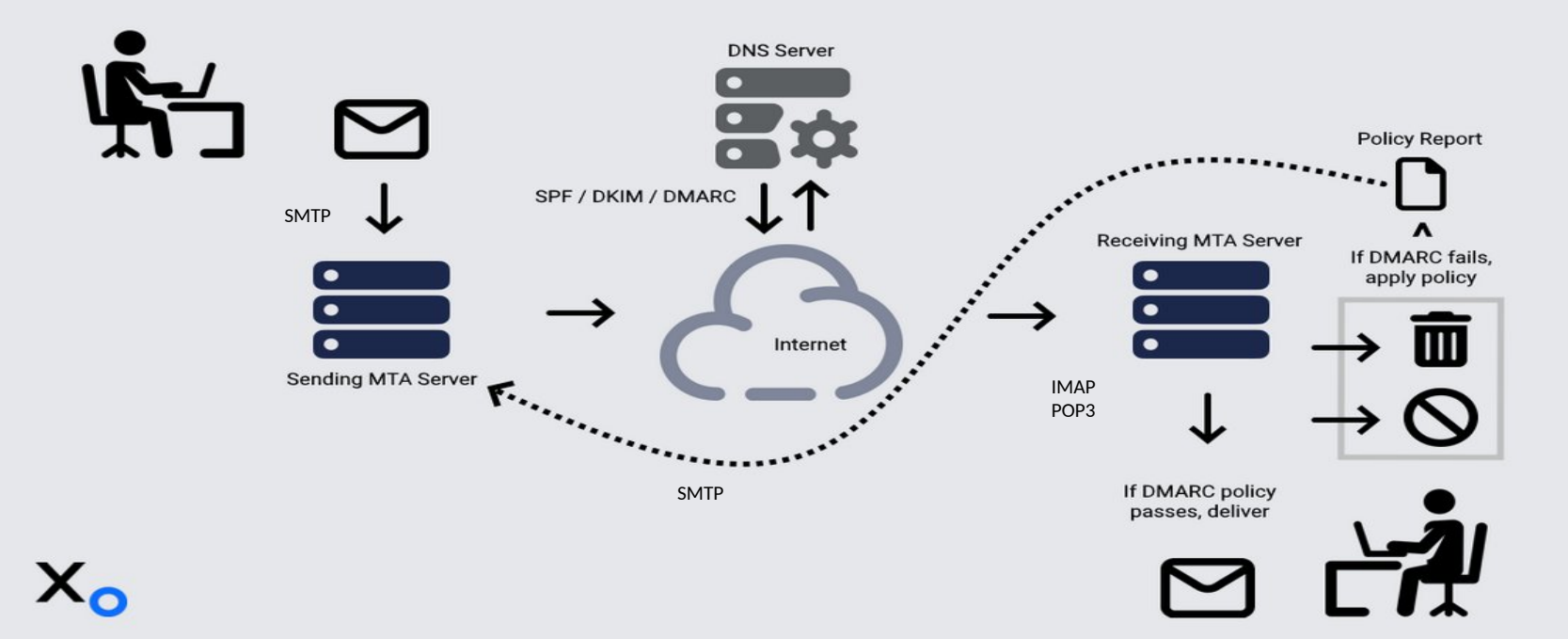


Figure1 : Schéma de principe de la messagerie électronique [***XoMEDIA, A Deep into Email Deliverability in 2024***]

1. Contexte

téicée® la raison de votre confiance

GREYC Electronics and Computer Science Laboratory

➢ **Les attaques de Phishing menacent les Organisations**

Données récentes

- **~91% de toutes les cyberattaques** sont les attaques de Phishing par mail [*TEHTRIS, 2024*]
- **~50% des grandes organisations** reçoivent en moyenne **5 e-mails de phishing ciblé par jour** [*TEHTRIS,2024*]
- Pour une entreprise, le coût moyen d'une attaque de phishing par mail réussie est d'**environ 4,65M USD** [*L.Ana, 2024*]

Les attaques de phishing sont rentables et, avec le temps, les attaquants affinent leurs méthodes pour gagner en crédibilité et augmenter leur nombre de victimes.

Figure2 : Evolution des attaques de phishing **[N.Quang & Selamat ; april 2022]**

1. Contexte

➢ Définition d'une attaque de Phishing

- *Phreaking* : piratage de ligne téléphonique
- *Fishing* : pêche
- Contraction de *fishing* et *phreaking*

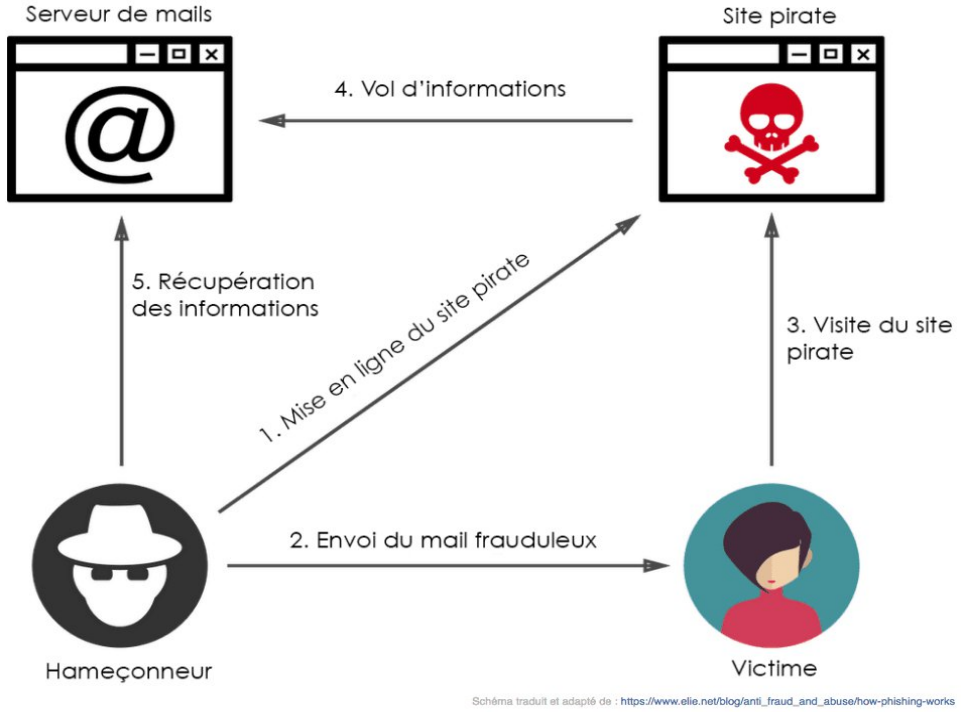


Figure3 : Synoptique d'une attaque de phishing par mail [**N.Quang & Selamat ; april 2022**]

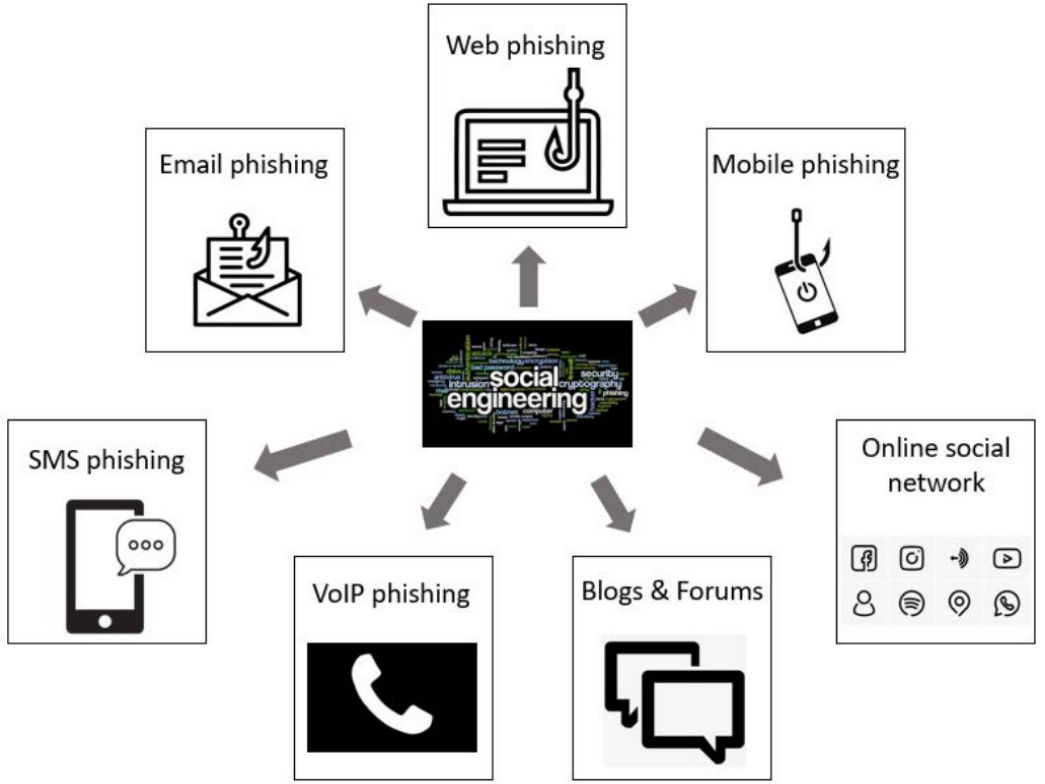


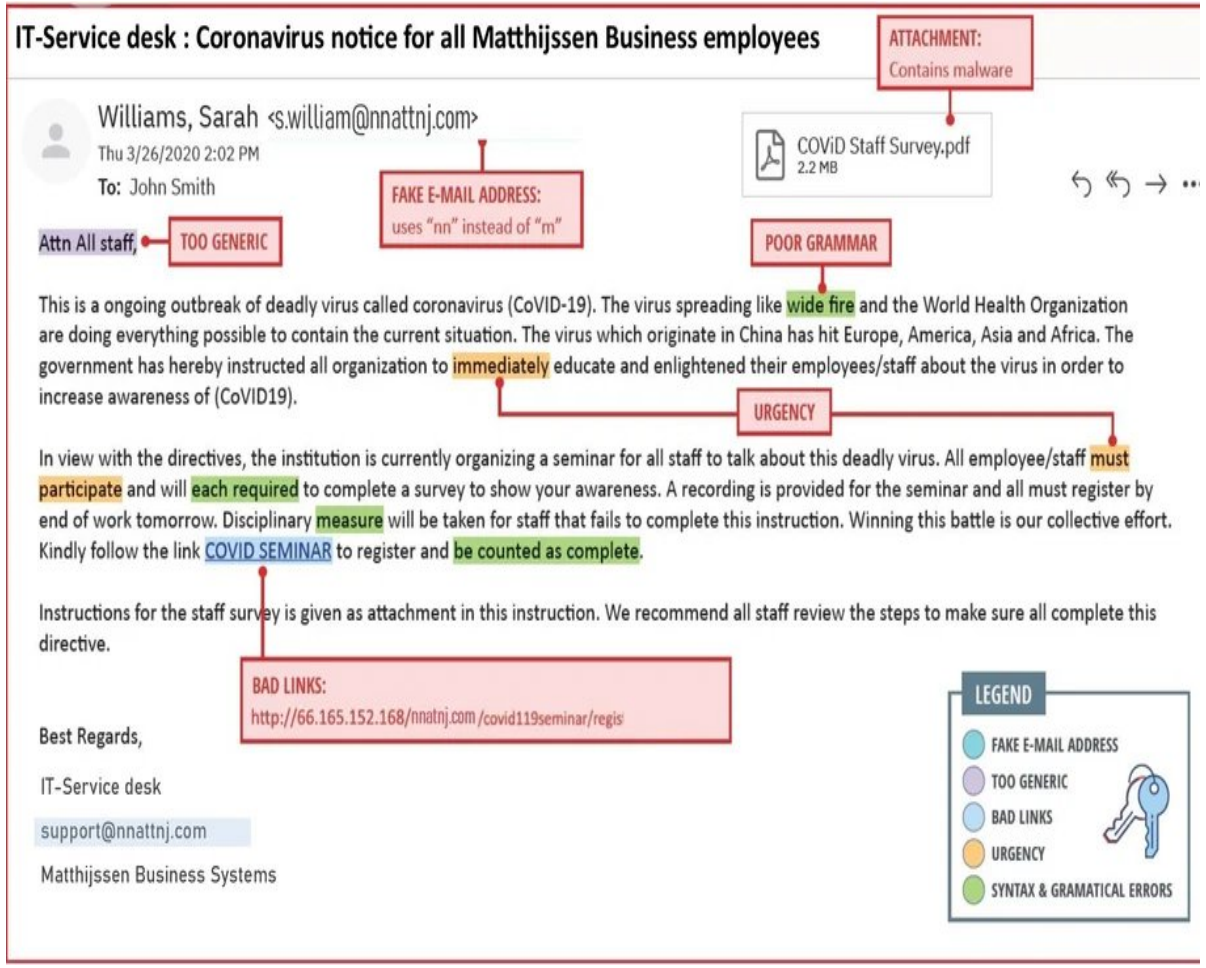
Figure4 : Techniques d'ingénieries sociales utilisées par les cyberattaquants pour lancer les attaques de phishing **N.Quang & Selamat ; april 2022**]

1. Contexte

téicée® la raison de votre confiance

GREYC Electronics and Computer Science Laboratory

➢ Exemple n°1 de mail de phishing



➢ Exemple n°2 de mail de phishing

2. Objectifs

Concevoir un système de protection robuste et fiable contre les attaques de phishing avec la détection automatique et la modération par l'humain.

- Analyser les différents aspects des attaques de type phishing ;
- Suivi des statistiques et faire une veille technologique sur l'évolution des escroqueries par phishing ;
- Proposer des solutions comme contre-mesures en utilisant un faisceau d'indices ;
- Utilisation des techniques d'IA basées sur le ML, le DL couplées aux méthodes de TAL ;
- Approche collaborative exploitant le résultat de l'analyse automatique et le signalement par un utilisateur (modérateurs DSI et ou RSSI pour confirmer ou infirmer le statut du mail) ;

3. Enjeux Scientifiques

- Mise au point d'une méthode performante par apprentissage permettant de décider si un mail contient une attaque de phishing;
- Création de partage d'un dataset ;
- Mécanisme d'anonymisation/pseudonymisation l'analyse manuelle d'un mail suspect et pour création de partage d'un dataset ;
- Définition d'une signature de quasi-similarité de mails pour la généralisation de la détection de mail de phishing personnalisé

5. Plan de Travail

- Réaliser un état de l'art de toutes les facettes de la lutte contre le phishing;
- Créer une base de mails de phishing spécialisée sur le français ;
- Analyser l'impact de l'anonymisation et de la pseudonymisation sur la détection du phishing, les mesures de protection de la vie privée avec l'identification des données en terme de DCP sensible ainsi que la conception de services innovants sécurisés suivant la réglementation RGPD ;
- Analyser la structure et du contenu des mails de phishing,
- Génération d'empreintes de documents permettant une recherche approximative dans des espaces de grande dimension ;
- Utilisation des modèles de ML et de DL pour la prédiction et la détection des attaques de phishing, etc

6. État de l’art

Structure du mail, Format MIME, types de mails et indices exploitables

```
Date: Thu, 26 Sep 2024 12:43:16 +0200 (CEST)  1
From: Armand Florent Tsafack Piugie <armand-florent.tsafack-piugie@unicaen.fr>  2
To: ftsafack <ftsafack@teicee.com>,  3
    Armand Florent Tsafack Piugie <armand-florent.tsafack-piugie@ensicaen.fr>
Message-ID: <1208220690.4931428.1727347396944.JavaMail.zimbra@unicaen.fr>  4
Subject: =?utf-8?Q?TP_fouille_de_donn=C3=A9es_et_Appren?=  5
 =?utf-8?Q?tissage_Automatique_2A_INFO_TP2?=
MIME-Version: 1.0  6
Content-Type: multipart/alternative;  7
    boundary="=_bfe05c8f-4824-421a-b07a-f075983c82bc"
X-Originating-IP: [193.49.200.225]  8
X-Mailer: Zimbra 10.1.1_GA_4660 (ZimbraWebClient - FF115 (Linux)/10.1.1_GA_4660)
Thread-Index: zQ+7A2X0yDwTZyMc1k493GPGK7ID7w==
Thread-Topic: TP fouille de =?utf-8?Q?donn=C3=A9es?= et Apprentissage Automatique 2A INFO TP2
X-SPAM-LEVEL: Spam detection results:  0  9
    AWL                     -0.124 Adjusted score from AWL reputation of From: address
    BAYES_00                 -1.9 Bayes spam probability is 0 to 1%
    DKIM_SIGNED               0.1 Message has a DKIM or DK signature, not necessarily valid
    DKIM_VALID               -0.1 Message has at least one valid DKIM or DK signature
    DKIM_VALID_AU            -0.1 Message has a valid DKIM or DK signature from author's domain
    DKIM_VALID_EF            -0.1 Message has a valid DKIM or DK signature from envelope-from domain
    DMARC_PASS               -0.1 DMARC pass policy
    KAM_NUMSUBJECT            0.5 Subject ends in numbers excluding current years
    SPF_HELO_NONE           0.001 SPF: HELO does not publish an SPF Record
    SPF_PASS                -0.001 SPF: sender matches SPF record
    URIBL_DBL_BLOCKED_OPENDNS  0.001 ADMINISTRATOR NOTICE: The query to dbl.spamhaus.org was blocked due to
X-Bm-Milter-Handled: 20d7bb68-1eb8-43f6-8568-f3a57026b047
X-Bm-Transport-Timestamp: 1727347399696
X-BM-Event: 033bff5c-5a8f-408d-9516-3c48d0981b22; rsvp="true"

--=_bfe05c8f-4824-421a-b07a-f075983c82bc
Content-Type: text/plain; charset=utf-8  10
Content-Transfer-Encoding: quoted-printable  11

Nouvelle demande de r=C3=A9union ci-dessous=C2=A0:

Sujet=C2=A0: TP fouille de donn=C3=A9es et Apprentissage Automatique 2A IN=
FO TP2=20
Organisateur: "Armand Florent Tsafack Piugie" <armand-florent=2Etsafack-pi=
ugie@unicaen=2Efr>=20

Endroit=C2=A0: Salle E205=20
Heure: Mardi 17 D=C3=A9cembre 2024, 15:45:00 - 17:45:00 GMT +01:00 Bruxell=
es, Copenhague, Madrid, Paris
=20
Invit=C3=A9s: ftsafack@teicee=2Ecom; armand-florent=2Etsafack-piugie@ensic=
aen=2Efr=20


*~*~*~*~*~*~*~*~*


--=_bfe05c8f-4824-421a-b07a-f075983c82bc
Content-Type: text/calendar; charset=utf-8; method=REQUEST; name=meeting.ics  12
```

Figure 5 : Partie de l'enveloppe du code source d'un mail

- **Les types de mails**
  - **Les spams**
  - **Les scams**
  - **Les mails de Phishing**
  - **Les mails de spear phishing**
  - **Le Whaling ou BEC**
  - **Ransomwares ou rançongiciels**

6.1. Revue des Bases de Données des mails de phishing

Dataset	Size	Parties de mail utilisées	Classe 1	Classe 2	créateur	Lien téléchargement
Phishing Archive (APWG)		Body			Anti-Phishing Working Group	
Corpora of the SpamAssassin project	>9000	Body + Header	Easy_ham and hard_ham (Apache Software Foundation (2014) Spamassassin Public Corpus, 2006) Easy ham_2	Spam_3		https://spamassassin.apache.org/downloads.html
SpamAssasin containing and emails available online		Body	2551 ham	501 spam		https://www.kaggle.com/datasets/veleon/ham-and-spam-dataset
Enron	>500,000	Body	43 000 ham	50,000 spam	CALO Project (A Cognitive Assistant that Learns and Organizes)	***Enron all :*** http://www.cs.cmu.edu/~enron/enron_mail_20150507.tar.gz ***Enron-Spam Dataset*** ***- Lien Original :*** https://www2.aueb.gr/users/ion/data/enron-spam/ ***- Version Prétraité :*** https://github.com/MWiechmann/enron_spam_data/
Enron1	6447	Body	3228 ham	3219 spam emails		***https://www2.aueb.gr/users/ion/data/enron-spam***
TREC corpus		Body	39,399 legitimate emails	52,790 spam emails		**2007 TREC Public Spam Corpus** ***Version Originale :*** https://plg.uwaterloo.ca/~gvcormac/treccorpus07/ ***Version prétraité :*** https://www.kaggle.com/datasets/bayes2003/emails-for-spam-or-ham-classification-trec-2007
CSDMC 2010 SPAM	4,327 training emails	Body + Header	2,949 ham	1378 spam	Organizers of data mining competition	https://github.com/jdwilson4/Intro-to-Machine-Learning/tree/master/Data/SPAMData
Nazario/Phishing Corpus				7315 spam	Jose Nazario	https://monkey.org/~jose/phishing/
SMS Spam Collection	5572		4825 ham	747 spam	University of California irvine	https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset
IWSPA-AP-2018	10,306	Body + Header	9174 ham	1132 spam	Organizers of IWSPA 2018 competition	https://dasavisha.github.io/IWSPA-sharedtask/
live emails received by WestPac and their customer	659,673 emails		613,048 emails are legitimate	46,525 of the emails are phishing emails		

6.2. Deep Learning Models for Phishing Email Detection

Article Title	year	Authors	Purpose	Source de Données	Parts of Email used	Taxonomy of feature extraction used	Models	Precision	Accuracy	Score f1
Advancing Phishing Email Detection : A Comparative Study of DL Models	March 2024	Najwa Altwijry, Isra Al-Turaiki, Reem Alotaibi and Fatimat Alakeel	Text Classification	Phishing Corpus (7315 mail de phishing) and Spam Assassin (6047 email samples, in which there are 1897 spam emails and 4150 legitimate emails)	Body and Subject	GloVe	1D-CNNPD with recurrent layers namely LSTM, Bi-LSTM, GRU and Bi-GRU		98.87%, 99.23%, 99.34%, 99.01%, 99.68%,	-, 99.20%, 99.31%, 99.66%, 99.66%
Phishing Email Detection Model Using Deep Learning	October 2023	Samer Atawneh and Hamzah Aljehani	Phishing Email Detection	**Enron, SpamAssassin** et **UCI** pour constituer un jeu de données de mails étiquetés (**en mails légitimes et mails de phishing**) de **18040 échantillons**	Body	Tokenization (Keras tonkenizer class for first 3 models and Hugging face tokenizer class for BERT wich use the WordPiece method)	CNN, RNN, LSTM, BERT		98.74%, 98.58%, 98.89%, 99.61%	
Phishing Email Detection using improved RCNN with multilevel vectors and attention mechanism	March 2019	Yong Fang , Cheng Zhang, Cheng Huang , Liang liu, and Yue Yang	Classifying Phishing Email	First Security and Privacy Analytics Anti-Phishing Shared Task (IWSPA-AP 2018) : legitimate emails (Wikileaks), phishing emails (IT departments of university and Nazario Phishing corpora)	Subject and Body	Word2Vect	THEMIS that utilized an improved recurrent convolutional neural network (RCNN) model with multilevel vectors and attention mechanisms		99.848% and a low false-positive rate of 0.043%	
Phishing and Fraudulent Email Detection through Transfer Learning using pretrained transformer models.	November 2022	Bronjon Gogoi; Tasiruddin Ahmed	Phishing email Classification				BERT and DistilBERT		99%	

6.3. Machine Learning Models for Phishing Email Detection

Article Title	year	Authors	Purpose	Data Sources	Parts of Email used	Taxonomy of feature extraction used	Models	Precision	Accuracy	Score f1
Detecting Phishing Email Using Hybrid features	2009	Liping Ma,Bahadorrezda Ofoghi, Paul Watters, Simon Brown	Classifying phishing Email using Hybrid Features	WestPac emails	Subject, Body	Links, nonv links,nonmatching urls, forms, scripts, body BL words, subject BL wors, Information Gain	C4.5		99%	
Phishing Email Detection Technique by using Hybrid Features	2015	Isredza Rahmi A. Hamid and Jemal Abawajy	Binary classification problem	Phishing Corpus and SpamAssassin (pour constitué un small training set of 1000 emails)		Collection of nine structure- and behavior-based features : DES, SBW, URLD, URLS,URLIP,US,UD,DMID	SVM		97.25%	
Accurate spear phishing campaign attribution and early detection	2016	YuFei Han	semisupervised learning approach with a dataset of 1467 spearphishing emails and 4043 legitimate emails	Symantec's enterprise email scanning service	origin, text, attachment, and recipient features		K-nearest neighbor attribute graph to detect spear phishing attacks			90% and an FPR of 0.1 for known campaigns
Spam Email Detection Using ML Techniques	2023	Ioannis Moutafis, Antonios Andreatos and Petros Stefaneas	text classification	Spam Assassin, Enron1, SMS-Spam Collection	Body		SVM, k-nearest neighbor, naïve Bayes, neural network, recurrent neural network, AdaBoost, random forest, gradient boosting, logistic regression, and decision tree methods		For the first dataset, the best performance was 99.51%, achieved by the NN, while the SVM achieved 99.38% for the second dataset.	

7- Mes Travaux au cours de ce 1er Semestre

7- Mes Travaux au cours de ce 1er Semestre

téicée® la raison de votre confiance

GREYC Electronics and Computer Science Laboratory

- Entraînement de 11 modèles de ML et 1 modèle Deep sur la base constitué à partir des sources Enron et TREC 2007 *(~83k échantillons étiquetés en spam et ham)*

Tableau récapitulatif des performances des modèles de ML et DL

Algorithme	Accuracy	Score f1	Precision
ETC	**0.986**	**0.99**	**0.99**
RF	0.984	0.982	0.980
LR	0.981	0.98	0.974
BgC	0.972	0.97	0.97
MNB	0.955	0.96	0.955
GNB	0.960	0.96	0.957
AdaBoost	0.947	0.95	0.95
SVC	0.932	0.93	0.931
KN	0.915	0.92	0.867
DT	0.859	0.86	0.801
NN	0.971	-	-
DistilBERT	0.95	0.95	-

Tableau récapitulatif des performances des modèles de ML réalisé par (***Ioannis et al.,2023)*** sur 03 Datasets

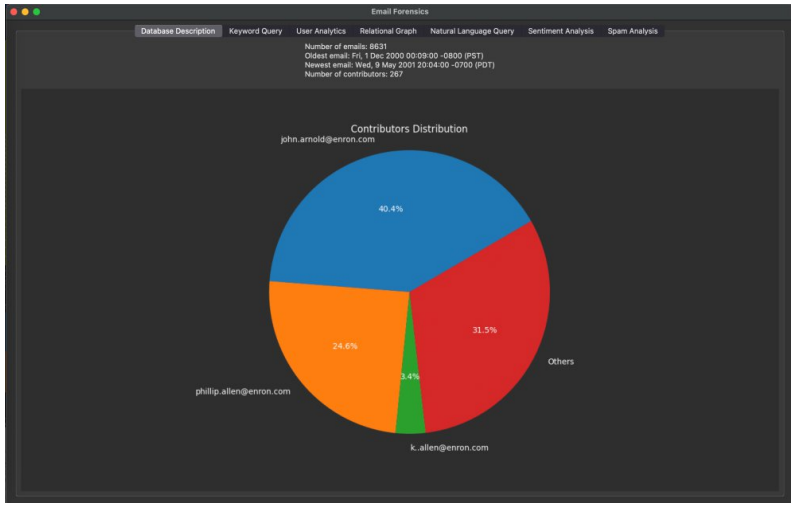
Algo/Dataset	SpamAssassin	Enron1	csv by F.Qureshi
SVM	**0.9918**	**0.9938**	**0.9839**
KNN	0.959	0.9853	0.9551
NB	0.9131	0.983	0.9543
NN	0.9951	0.9922	0.9823
RNN	0.9836	0.993	0.983
AB	0.982	0.983	0.9838
RF	0.982	0.972	0.9811
GB	0.959	0.966	0.959
LR	0.9541	0.9922	0.965
DT	0.9573	0.965	0.9786

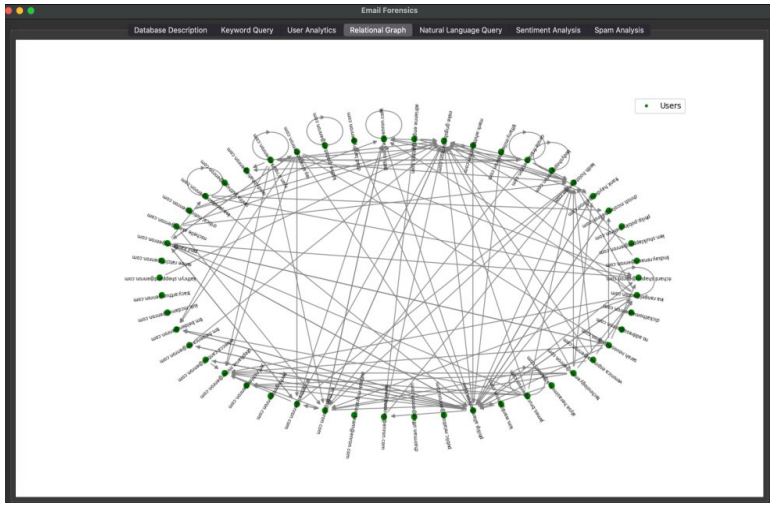
7- Mes Travaux au cours de ce 1er Semestre

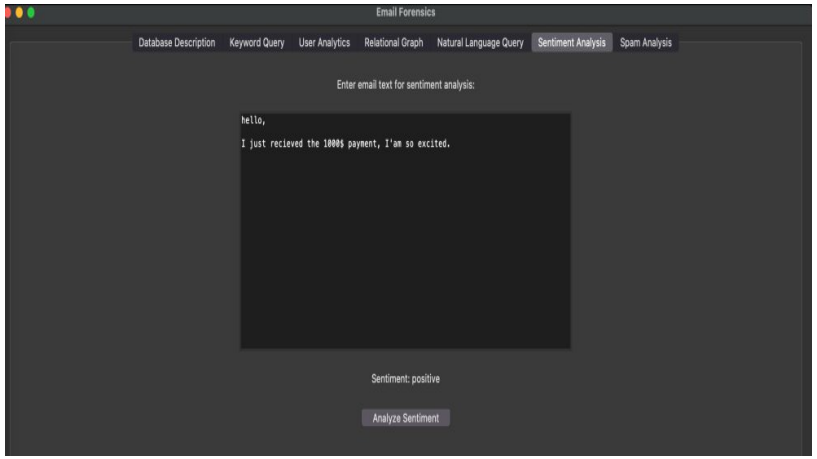
téïcée® la raison de votre confiance

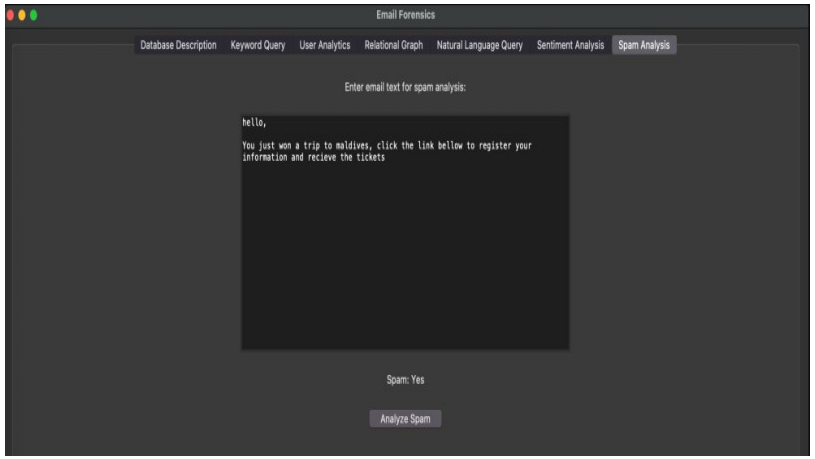
GREYC Electronics and Computer Science Laboratory

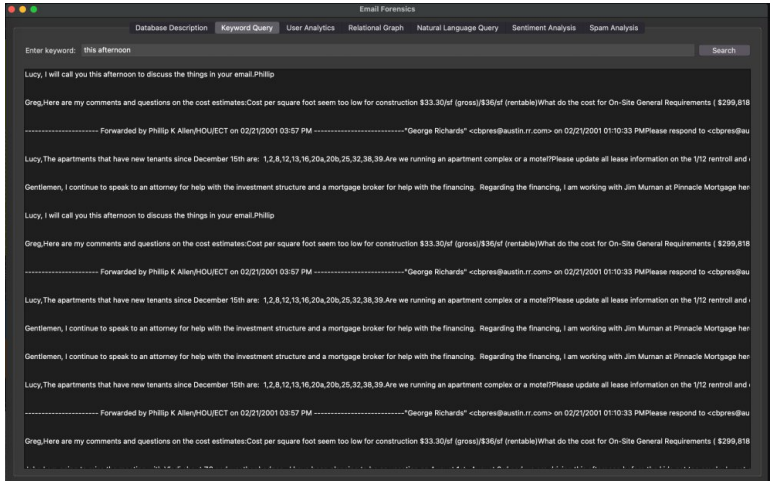
- École d'été sur la cybersécurité organisée par l'**ENSICAEN** : J'ai travaillé sur le projet de téïcée notamment sur l'***analyse forensique de la base de données d'Enron***











8- Formations suivies

Formations SYGAL

- Logiciel d'Apprentissage Automatique (**10h**)
- Comment faire confiance à l'Intelligence Artificielle : une introduction à l'IA de confiance (**03h**)
- Travail collaboratif et gestion de versions avec Git (**06h**)
- Comment faire confiance à l'Intelligence Artificielle : les différents ingrédients de l'IA de confiance (**04h**)

Formation ANRT

- Horizon Europe pour les jeunes chercheurs : A la découverte de opportunités de financement et des politiques de recherche **(4h)**

Conclusion

Tout comme les attaques de phishing par mails sont de plus en plus rentables (puisque les attaquants affinent leurs méthodes avec les techniques d'ingénieries sociales), les solutions de détections sont aussi croissantes (de part les architectures développées et entraînées dans la revue). Ce pendant, la difficulté d'avoir des jeux de données étiquetés et sans biais limite la capacité de généralisation des solutions existantes. D'où la pertinence et les enjeux de cette thèse.

Merci !